The KnowledgeMap Project: Development of a Concept-Based Medical School Curriculum Database

Joshua C. Denny, MD[1,2], Plomarz R. Irani[1,2], Firas H. Wehbe, MD[1], Jeffrey D. Smithers, MD[1,2], Anderson Spickard, III, MD, MS[1,3]

[1]Department of Biomedical Informatics, [2]Vanderbilt University School of Medicine, [3]Vanderbilt Department of Medicine – Vanderbilt University Medical Center, Nashville, TN

We developed the KnowledgeMap (KM) system as an online, concept-based database of medical school curriculum documents. It uses the KM concept indexer to map full-text documents and match search queries to concepts in the Unified Medical Language System (UMLS). In this paper, we describe the design of KM and report the first seven months of its implementation into a medical school. Despite being emphasized in only two first year courses and one fourth year course, students from all four classes used KM to search and browse documents. All faculty members involved with courses piloting KM used the system to upload and manage lecture documents. Currently, we are working with eight course directors to transition their courses to KM for next year.

INTRODUCTION

Medical educators have recognized the need for centralized access to curriculum documents[1-3] and more efficient accounting of what material has been taught.[4,5] Medical and dental schools have employed web-based curriculum databases[6,7] and indexed documents to UMLS concepts[8,9] for improved retrieval. Manual-entry databases can ease the time-intensive process of curriculum review and revision.[3] Early attempts to automate concept indexing of medical school documents produced suboptimal results.[10] Improvements in concept indexing may better equip administrators to identify gaps and overlaps in curricula and allow students and teachers to meet personal information needs. In addition to more accurate and powerful searches, a system mapping documents to UMLS concepts could leverage the semantic information provided in the UMLS.

The KnowledgeMap (KM) concept identifier was developed to extract concepts represented in medical educational texts. Initial analysis has shown that the KM concept identifier performed favorably compared to the National Library of Medicine's MetaMap[11] using selected subsets of curriculum documents.[12] This paper describes the system architecture and initial implementation of the KM concept indexer into a medical school curriculum database at Vanderbilt University School of Medicine.

SYSTEM DESIGN

Hardware and software

We implemented KM on three Microsoft® Windows®-based systems. We selected the Apache 2.0 series with OpenSSL as our web server and MySQL 3.23.51 as our database engine. All CGI scripts were written in the Perl programming language. Perl, Microsoft® VisualC++®, and Microsoft® VisualBasic® were used to create additional server software for KM.

Concept Identifier Algorithm

The KM Concept Identifier (KMCI) has been described elsewhere[11]; a brief overview is provided below. KMCI uses approximate natural language processing techniques and scoring heuristics to map noun phrases in text documents to concepts in the UMLS.[13] KMCI currently uses the 2001 edition UMLS.

The KMCI processes documents in three phases:

1. Sentence Identification and Normalization: First, KMCI identifies sentences within documents and removes outline headers, attempting to distinguish between true outline headers and meaningful abbreviations (such as "P. aeruginosa" or "IV" meaning "intravenous"). It then normalizes words and determines part-of-speech using a library from Cogilex, R & D, Inc.[14]
2. Concept Identification and NLP techniques: KMCI first selects a list of candidate UMLS concepts for each "simple" noun phrase – those noun phrases consisting only of nouns and/or adjectives and associated modifying adverbs and numbers. If a match is not found, KMCI generates semantic and derivational variants for each word. If KMCI finds a set of candidate concepts that matches the concept, it then attempts combinatory matching with other noun phrases linked by conjunctions, prepositions, or linking verbs. During this phase, it also attempts to distribute modifying adjectives (translating "small and large intestine" into "small intestine and large

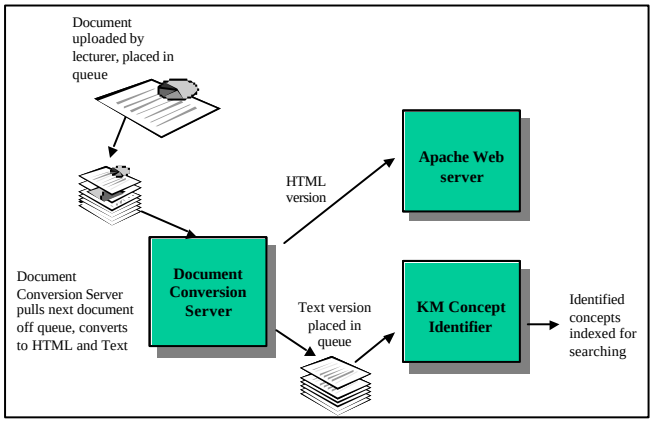


Figure 1. Automated processing of uploaded documents.

intestine") to more accurately represent these concepts.

3. Concept disambiguation: Scoring of candidate concepts occurs on phrase, context, and document levels. KMCI prefers candidate concepts that most closely match the document phrase. During document processing, KMCI creates a list of "exactly-matched" concepts. Candidate concepts are scored based on similarity to other exactly-matched concepts. KMCI also favors candidate concepts that co-occur with exactly-matched concepts in Medline abstracts.

Document Corpus Processing

Prior to development of the KM interface, we collected 571 documents from the first two years of medical school. Since these documents were collected from previous curricula, they were termed "legacy" documents. When a user uploads a new document covering the same material as an existing legacy document, we replace the legacy document with this newer version.

Faculty members populate the document database by uploading HTML, Microsoft® Word®, or PowerPoint® documents for automatic processing (see **Figure 1**). The document conversion server uses Microsoft Word and PowerPoint to convert these documents to HTML and ASCII text. Since PowerPoint only allows one to save the outline text of a slide (ignoring textboxes and tables), we developed our own program to extract all text from PowerPoint slides, including textboxes and tables. KM indexes all documents by both concept and normalized word using the KMCI.

Concept-based Navigation Tools

We built three UMLS Metathesaurus-based tools to navigate through documents: a concept-based search, a tool to generate relevant PubMed queries based on a lecture, and a "content coverage" query.

Search. A user can submit a Boolean query via the web interface. Each component of the Boolean query is submitted individually to a concept server based on the KMCI (**Figure 2**), except that it returns multiple concepts for ambiguous matches. For example, a search for "CHF" returns "congestive heart failure," "congenital hepatic fibrosis," and "Crimean hemorrhagic fever." KM then searches the index of concepts from the document corpus and presents a list of documents that match the concept query, ordered by course and relevancy score. If a search by concept returns no documents, then KM normalizes each word of the query and returns the documents matching the normalized words in the search term.

Relevant PubMed query. Users may perform a PubMed query to find relevant Medline articles to a viewed document. The PubMed query is constructed from the document title (ignoring common words such as "clinical correlation," "lecture," or "presentation") and the most frequent MeSH concepts in the document. These parameters are adjusted to yield between 1 and 200 Medline search results.

Content Coverage query. This option, currently available only to course directors and administrators, allows a user to search for a "metaconcept." Examples include a search for "genetics" or "women's health." After a user submits a metaconcept or area of interest, KMCI identifies the metaconcept. KM constructs a list of child and child-like concepts (as defined by the relationships in the Metathesaurus) to be included in the search.[15] For example, for the metaconcepts "genetics," KM constructs a list of "subconcepts" such as "DNA repair," "exon," and "genotype." The user can select all or a set of concepts to describe the metaconcept.

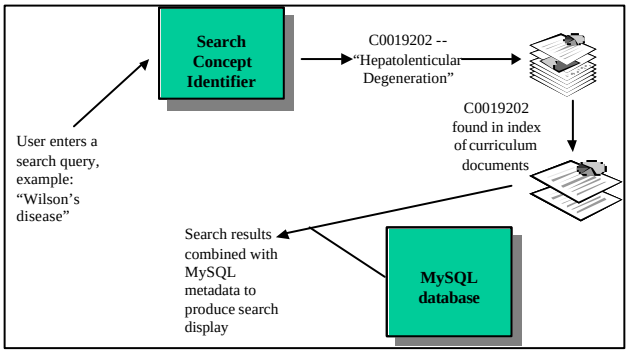
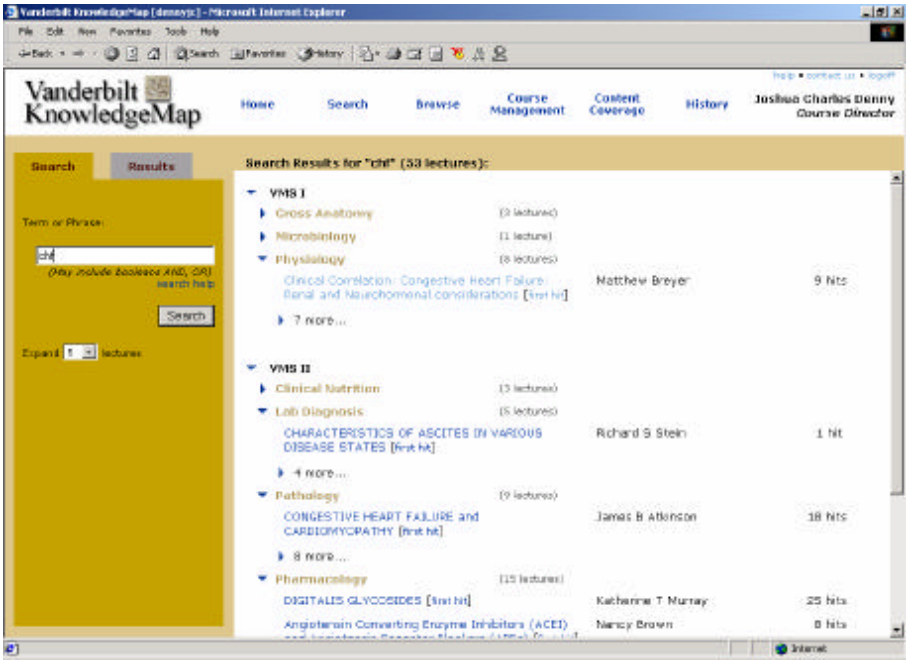


Figure 2. Search by concept. If a search by concept fails, KM repeats the process using normalized words.



session. For documents, this database only stores metadata. The documents are stored via the file system with unique numbers that are keys in the relational database. KM also maintains an independent database of the concepts and normalized words in each document for searching, concept coverage queries, and PubMed queries.

Interface

We restrict KM login to active Vanderbilt medical students and academic faculty via an SSL-encrypted campus-wide authentication procedure. In response to intellectual property concerns, users cannot "bookmark" documents for immediate access; all KM usage requires login.

KM then identifies all curriculum documents containing this set of concepts, sorting them by relevance.

Database design

A relational database, implemented in MySQL, supports the course and document management functionality of KM. We used an entity-relationship model[4] to represent such entities as "document," "person," "class session" (i.e., a lecture), and "course." The entity-relation schema was mapped to a relational model to allow for many-to-one and many-to-many relationships, such as multiple documents per session and multiple lecturers per

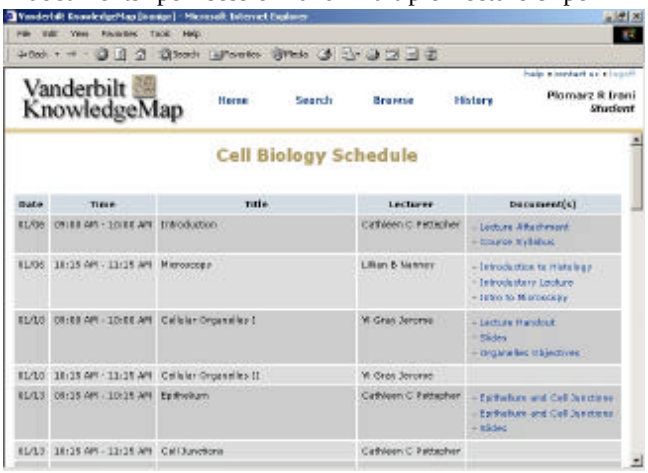


Figure 4. Course home page view.

We designed the KM interface to provide rapid access to course documents and searching functions. Upon login to the system, the search panel immediately appears. A toolbar across the top of the screen allows rapid access to the search, browse, course management, content coverage, and search history (see **Figure 3**). KM displays search results by year and course in a collapsible outline format, with the most relevant documents automatically expanded. Users may browse the curriculum to find a particular lecture or document displayed by title, lecturer, and date within a course. Users can also select a course "home page" view that displays all sessions in a course and its documents (see **Figure 4**).

Course directors use the course management interface to modify their course schedules and assign lecturers to sessions. Faculty members use the course management interface to view, delete, or replace documents uploaded to their sessions.

METHODS

Pilot design

Three courses piloted KM: two required first year courses (Gross Anatomy in the fall and Cell Biology in the spring) and one elective fourth year course (Clinical Management). In addition, four documents

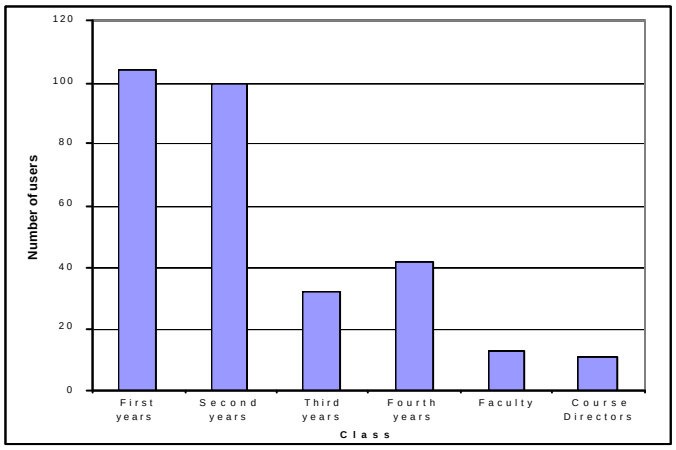


Figure 5. Number of users who accessed KM. "First years" through "Fourth years" refers to that class of Vanderbilt medical students.

of the second year neuroanatomy course were displayed on KM. No documents of the third year classes were placed online.

Analysis

To measure system usage, KM logs user interactions with the site. We analyzed the KM log file for usage statistics of the different components of KM (e.g., search, browse), comparing it with the log file generated by the Apache web server. We calculated a "search/browse ratio" for each medical school class by dividing the number of documents viewed via a search by the number of documents viewed via a browse. A "browse" included documents accessed via a course home page or via the "browse" function on the toolbar. We eliminated the contributions of the developers from the total counts and calculations. All statistical comparisons were made using one-way ANOVA tests with Stata 7.0 (Stata Corporation, College Station, Texas).

RESULTS

During the period from 8/20/02 through 3/5/03, 317 unique users logged in to KM 4,639 times. Total server downtime was 20 hours, a little over half of which was for scheduled upgrades or bug fixes. Students accounted for 93% of the login traffic; 74% was from first year students. Of the 884 total active documents available (legacy and uploaded), 665 documents (75% of those available) were viewed a total of 8,571 times. There were 820 searches and 51 content coverage queries. The eight faculty members involved in the two first year courses uploaded 252 documents. The authors uploaded 101 documents, 81 of which were for the fourth year clinical management course. Six additional faculty members and nine additional course directors also accessed KM. See also **Figures 5** and **6**.

DISCUSSION

We piloted KM with two first year courses and one fourth year course; interestingly, login activity and searches extended to each class year and involved over 300 legacy documents not uploaded as part of the courses officially piloted with KM.

Most of the activity by the first and second year students involved browsing, whereas in later years, searching became more prominent. This is not unexpected, since first year students typically use KM to browse documents while third and fourth year students likely use KM to answer clinical queries. In addition, lecturers would not have a need to browse documents but would instead be more likely to search for previously covered topics. We expect that the number of searches will increase as more courses add content to KM and as subsequent classes use the system more.

Limitations caution the interpretation of these results. These results show usage data, not satisfaction or accuracy. We have not validated our searching algorithms in real-life testing nor compared them to more simple word-based search algorithms. Given that the search/browse ratio was calculated based on the number of documents viewed via each interface, the difference in ratios between classes could be affected by the users' familiarity with the system and curriculum – first year students could be more likely to find their target document in a search with fewer attempts. It is unknown what effect additional

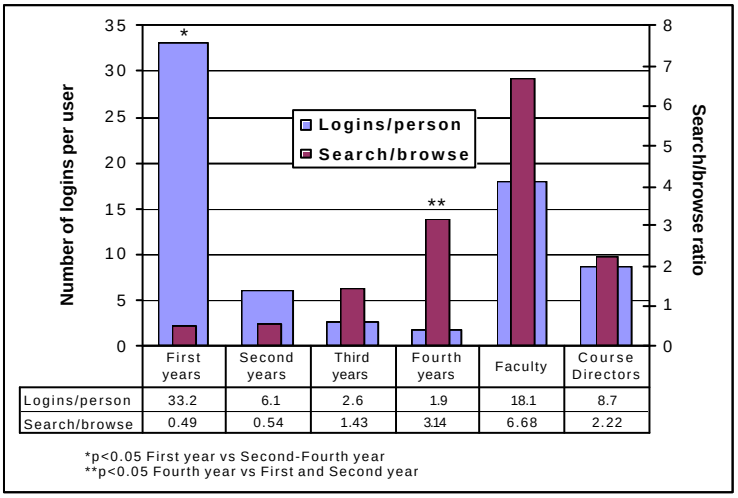


Figure 6. Logins and search/browse ratios per user.

courses will have on searching behavior. The finding of our server downtime of 20 hours may underestimate its impact: much of this downtime occurred during peak usage, including one night preceding an exam. (Subsequent versions of Apache have improved the system's stability.)

Evaluation of the first seven months of use of KM provides directions for future improvements. We will now include document metadata in searches to allow for searches for author or course names not included in the document. We will improve the PubMed query tool by determining those "key" concepts of a document instead of merely the most frequent MeSH concepts in a lecture, or from determining which concepts co-occur together.[16] In its present form, the content coverage query can easily result in hundreds of documents. A partitioning algorithm to identify groups of "high coverage" and "low coverage" will improve its usability. Other suggestions from end-users will be included, such as providing links to other documents (or journal papers) and searches of outside resources, such as medical dictionaries.

Currently, most courses at Vanderbilt rely heavily on slide presentations and paper handouts. Students collect these in notebooks, forming an out-of-date and unwieldy reference tool during their clinical training. The most important impact of KM may be in creating a more "electronic" atmosphere that promotes easier access, sharing, and integration of material. Future evaluation of KM should measure attitudes and assess behavioral changes based on its use, including collaboration between professors.

The Dean of the Medical School has voiced support for KM and is providing financial resources to sustain a programming team to maintain and innovate KM. Currently, we are working with eight course directors to assist them in transitioning their courses to KM. The true test of KM's usefulness will be in sustained use for content delivery and evaluation.

ACKNOWLEDGEMENTS

We would like to thank Art Dalley, Ph.D., and Cathleen Pettepher, Ph.D., for piloting their courses on KM. We would also like to thank Cogilex, R & D, Inc for providing their part-of-speech tagging software free of charge.

REFERENCES

1. Hamilton J. McGill makes Canada's first attempt to put medical curriculum in computerized format. CMAJ 1996 Jun 1;154(11):1731-2
2. Mattern WD, Anderson MB, Aune KC, Carter DE, Friedman CP, Kappelman MM, O'Connell MT. Computer databases of medical school curricula. Acad Med 1992 Jan;67(1):12-6
3. Cohen JJ. CurrMIT: you've gotta use this thing! Acad Med 2000 Apr;75(4):319
4. Kanter SL. Information management of a medical school educational program: a state-of-the-art application. J Am Med Inform Assoc, 1996; 3:103-111.
5. Report I: Learning objectives for medical student education: Guidelines for medical schools, Medical School Objectives Project, January 1998, American Assoc of Medical Colleges, http://www.aamc.org/meded/msop/msop1.pdf. Accessed 3/12/03.
6. Ward JP, Gordon J, Field MJ, Lehmann HP. Communication and information technology in medical education. Lancet 2001 Mar 10: 357(9258): 792-6
7. Zucker J, Chase H, Molholt P, Bean C, Kahn RM. A comprehensive strategy for designing a Web-based medical curriculum. Proc AMIA Symp. 1996:41-
8. Kanter SL. Using the UMLS to represent medical curriculum content. Proc Annu Symp Comput Appl Med Care. 1993:762-5.
9. Lee MY, Albright SA, Alkasab T, Damassa DA, Wang PJ, Eaton EK. Tufts Health Sciences Database: lessons, issues, and opportunities. Acad Med. 2003 Mar;78(3):254-64.
10. Kanter SL, Miller RA, Tan M, Schwartz J. Using POSTDOC to recognize biomedical concepts in medical school curricular documents. Bull Med Libr Assoc. 1994;82(3):283-7.
11. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21.
12. Denny JC, Smithers JD, Miller RA, Spickard-III A. "Understanding" medical school curriculum using KnowledgeMap. J Am Med Inform Assoc, 2003: 2003 Mar 28 [Epub ahead of print].
13. National Library of Medicine. UMLS Knowledge Sources, 12th Edition, 2001.
14. Cogilex R & D, Inc. Available at: http://www.cogilex.com
15. Smithers JD, Denny JC, Spickard-III A, Miller RA. Using concept markers to find genetics content in a medical school curriculum. Proc AMIA Annu Fall Symp 2002: 1167.
16. Miller RA, Gieszczykiewicz FM, Vries JK and Cooper GF. CHARTLINE: Providing bibliographic references relevant to patient charts using the UMLS Metathesaurus knowledge sources. Proc 16th SCAMC. 1992:86-90.